ELSEVIER Nuclear Instruments and Methods in Physics Research A 440 (2000) 202–212

NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A

www.elsevier.nl/locate/nima

Experience with fibre channel in the environment of the ATLAS DAQ prototype " $-$ 1" project

G. Ambrosini[a], H.P. Beck[b], D. Francis[a], M. Joos[a], G. Lehmann[a,b], A. Mailov[c], L. Mapelli[a,1], G. Mornacchi[a], M. Niculescu[a,d], K. Nurdan[c], J. Petersen[a], D. Prigent[a], J. Rochez[a], M. Romano[e], R. Spiwoks[a,*], L. Tremblet[a], G. Unel[c], E. van der Bij[a], T. Wildish[a]

[a] *CERN, EP Division, CH-1211 Geneva, Switzerland*
[b] *Laboratory for High Energy Physics, University of Bern, Switzerland*
[c] *Bogazici University, Istanbul, Turkey*
[d] *Institute of Atomic Physics, Bucharest, Romania*
[e] *Politecnico di Milano, Italy*

Received 12 May 1999; accepted 10 July 1999

Abstract

Fibre Channel equipment has been evaluated in the environment of the ATLAS DAQ prototype " $-$ 1". Fibre Channel PCI and PMC cards have been tested on PowerPC-based VME processor boards running LynxOS and on Pentium-based personal computers running Windows NT. The performance in terms of overhead and bandwidth has been measured in point-to-point, arbitrated loop and fabric configuration with a Fibre Channel switch. The possible use of the equipment for event building in the ATLAS DAQ prototype " $-$ 1" has been studied. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Fibre channel; Event building; Data acquisition

1. Introduction

The data acquisition (DAQ) system of the ATLAS experiment [1] at the LHC is expected to contain an event building (EB) system with total bandwidth of several Gbyte/s. This system shall be able to assemble event fragments from a few hundred to thousand sources at an input frequency of a few kHz. A list of the parameters is shown in Table 1. The Fibre Channel (FC) standard is a high-speed data transfer standard which offers both the high bandwidth and the high degree of connectivity, that is required for this system.

The final design for the ATLAS DAQ system is not scheduled to take place before 1999 and further investigations of detector requirements, hardware and software technologies as well as integration issues are still needed. The ATLAS DAQ group is approaching these predesign studies with a structured prototype [2] which will be based on the

* Corresponding author. Tel.: +41-22-767-3871; fax: +41-22-782-4897.

E-mail address: Ralf.Spiwoks@cern.ch (R. Spiwoks)

[1] ATLAS DAQ Prototype " $-$ 1" Project leader.

PII: S0168-9002(99)00863-3

Table 1
Parameters for the ATLAS event building system

Event size per source	1–15 kbyte	# Sources	100–1500
Input frequency	1–5 kHz	# Destinations	100–200
Data rate per source	1–75 Mbyte/s		
Total data rate	1–20 Gbyte/s	# Nodes	200–1700

current understanding of the ATLAS DAQ architecture [1]. The prototype (called " − 1") will contain a fully functional small-scale EB prototype with around ten sources and ten destinations. This prototype shall assemble event fragments, of 1–10 kbyte, from all sources in one destination, and several events may be assembled concurrently in several destinations. Performance is not the primary issue for this prototype but shall be watched in comparison to the requirements for the final system.

FC equipment has been investigated in order to learn about this technology and to understand the feasibility of FC for the EB prototype. Several evaluation studies of FC cards in point-to-point, arbitrated loop and fabric configuration have been carried out. The results will be presented in this paper.

2. System overview

2.1. Fibre channel standard

The Fibre Channel standard [3] defines a high-speed data transfer mechanism that can be used for networking, storage and data transfer. Fibre Channel is organized into the following levels:

- FC-0 defines the physical media, like optical and copper media, for a variety of data rates and transfer distances.
- FC-1 defines the transmission protocol, the 8-bit/10-bit coding and the receiver and transmitter state descriptions.
- FC-2 defines the signalling protocol, including control primitives, the frame structure and different classes of service.
- FC-3 defines a set of services which are common among multiple ports of a node.
- FC-4 defines the mapping between the lower levels of FC and upper level protocols, e.g. the Small Computer System Interface (SCSI) and the Internet Protocol (IP).

FC supports several topologies: point-to-point, arbitrated loop and fabric. In a point-to-point topology two nodes are directly connected. In an arbitrated loop several nodes are connected on a shared medium. In a fabric topology ports are connected to a node in the fabric (i.e. a switch or a network of switches) which routes data traffic from one port to another.

Data are organized in sequences which are made from frames with a length of up to 2112 byte. Frames can be sent one at a time (single frame sequence, SFS) or in groups sharing the same header (multi-frame sequence, MFS).

Several classes of service are defined: class 1 is a service for transfers with dedicated connections. Class 2 is a service for connection-less transfers where frames are acknowledged. Class 3 is a service for datagram tansfers.

Flow control is defined for classes 2 and 3 on a credit-based buffer-to-buffer (BB) flow control between every two directly connected ports: the receiver sends R_RDY primitives whenever it is ready to receive more frames. Classes 1 and 2 use a credit-based end-to-end (EE) flow control between two end-ports: the receiver sends ACK frames to acknowledge the reception of frames.

End-ports have to exchange service parameters, including BB and EE credit, before transferring data. This can be done implicitly by assuming parameters or by an initialization protocol using an

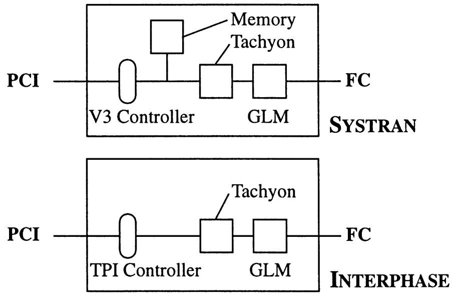

Fig. 1. FC Network Interface Cards.

PLOGI frame to an end-port or an FLOGI frame to a port on the fabric.

2.2. Network Interface cards

Two types of FC/PCI cards, from Systran Corp. and Interphase Corp., have been tested (see Fig. 1).

The FibreXpress FX cards from Systran Corp. [4] have a local memory into which data have to be moved before they are picked up by the FC protocol chip. The V3 controller is a bridge between PCI and the local memory bus. The cards are available in the PCI and the PMC form factor. The 5526 PCI FC adapters from Interphase Corp. [5] have a bridge (TPI) between PCI and the FC protocol chip. Both FC cards use the Tachyon chip from Hewlett-Packard Comp. [6] as the FC protocol chip and use multi-mode optical link modules (GLM) running at 1.0625 Gbit/s for the physical medium.[2] Whenever necessary PCI/PMC or PMC/PCI adapters from Technobox Corp. [7] have been used to adapt the cards from one form factor to the other.

The Tachyon chip handles several queues for input and output. The outbound command queue contains headers and data blocks for outgoing sequences. When the data are sent a message about the completion is put into the incoming message queue. The SFS and MFS queues contain data blocks for incoming data. When a complete sequence has been received a message about the reception is put into the incoming message queue.

2.3. Host computers

Two different host systems have been used for the evaluation of the FC equipment: RIO2 and PC. The RIO2 8060 or 8061 is a VME-based I/O board from CES [8] with a PowerPC running at 64 or 96 MHz, respectively. The PowerPC and its memory are connected to the PCI bus via the 27-82660 (Lanaï-Kauaï) PCI bridge from IBM [9]. The RIO2 has two PMC slots.[3] The operating system used on the RIO2 was LynxOS 2.4 [10], a real-time flavour of UNIX. The PC is based on a Pentium processor running at 166 MHz. The operating system used on the PC was Windows NT 4.0.

2.4. System software

For LynxOS a low-level library [11] was developed. This library maps the resources of the FC cards directly into the user space. The library allows to initialize the FC card, to send and receive sequences. On the Systran cards a send or receive consists of two transfers: the application moves data between the host memory and the card's memory using DMA, and the Tachyon chip moves the data between the card's memory and the FC link. On the Interphase cards the transfer is done in one step between the host memory and the FC link (see Fig. 1). The frame size can be programmed on the Tachyon chip and was chosen to be 2 kbyte for all tests. All messages are sent in one sequence (SFS or MFS, depending on the message size) and for every class 1 transfer a dedicated class 1 connection is set up before and removed after the sequence. The sending of several messages is pipelined using the Tachyon chip's outbound command queue. For the incoming message queue, polling was used. An interrupt-driven scheme was implemented and

[2] Some tests presented in this paper have been repeated with twin-ax copper link modules running at 1.0625 Gbit/s without any significant difference in the results.

[3] Some tests have been carried out on RTPC 8067 [8] which has the same architecture as the RIO2 8060 but only one PMC slot.

tested but not further investigated because it will not be used in the final application.

For Windows NT the FX light-weight protocol (FXLP) driver from Systran was used. The FXLP defines a simple, connection-oriented protocol in the FC-4 layer. Connections between two applications running on two different nodes are opened and closed using FC sequences. A blocking send and receive mechanism is used. Class 3 is used for the transfer of data which are sent and received in a sequential way, i.e. one at a time, doing DMA and FC transfer separately. FXLP is implemented for a loop topology, but a switch topology is not supported yet. Multicast and broadcast are not supported either. For tests between RIO2 and PC the LynxOS low-level library was extended to follow the definition of the light-weight protocol defined by FXLP. It implements the same API as for the Windows NT driver and represents a relatively small layer on top of the low-level library.

2.5. Method of measurements

The aim of the evaluation was to measure latency and throughput of transfers. Therefore, test programs consisting of a send application from one or several senders to one or several receivers were used. The sender applications run in a loop sending messages as fast as possible while the receivers run in a loop waiting for messages to arrive. The time for individual transfers of a given message size is measured in a loop over many transfers. The measurement is repeated for different sizes and a linear fit of transfer time versus message size is used to obtain *overhead* and *bandwidth* parameters as in

$$\text{time} = \textit{overhead} + \frac{\text{size}}{\textit{bandwidth}}. \tag{1}$$

The *overhead* is a measure for the transfer overhead resulting from hardware and software. For class 1 it also includes the overhead for establishing a class 1 connection, and for classes 1 and 2 it includes the round-trip-time because the ACKs have to be received before the transfer is finished. The error of the measurement for the *overhead* is about 1 μs. The *bandwidth* parameter is the saturation bandwidth for large messages and is mainly determined by the hardware (e.g. DMA). The error of the measurement for the *bandwidth* is about 0.2 Mbyte/s.

For detailed and more precise measurements an FC analyser has been used. The analyser uses the FC pre-processor from RMKI/KFKI [12] which consists of an optical splitter and a pre-processor for the 8-bit/10-bit coding. The output of the pre-processor can be examined with a Hewlett-Packard logical state analyser.

3. Performance in point-to-point configuration

3.1. Systran cards on RIO2s

Fig. 2 shows the data rate versus the size of the messages between two Systran cards on RIO2s in point-to-point configuration [13]. The fit parameters are very similar for all classes with an *overhead* of 25 μs and a *bandwidth* of 54 Mbyte/s.

The *bandwidth* of 54 Mbyte/s represents 54% of the FC capacity. This value is the result of two factors: the architecture where the concurrent DMA and the Tachyon transfers compete for the local memory bus, and the limited DMA band-

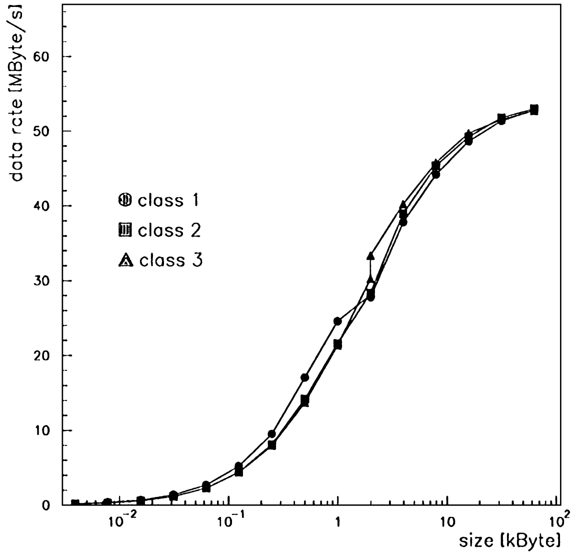

Fig. 2. Data rate for the Systran cards.

width itself. The concurrence on the local memory bus can be influenced by the Tachyon chip's read and write stream parameters. These parameters determine how long the Tachyon chip can hold the bus tenancy and thereby set the relative priority of the two transfers. The read and write stream parameters can have values of 1, 4, 16 and 64. The maximum data rate was found for both parameters having a value of 4. The small differences in the data rates for the different classes can be explained by the varying performance ratio of the two transfers for different sizes. The second limitation of the bandwidth is the DMA itself. This was measured separately to be 67 Mbyte/s reading from the host's memory and 58 Mbyte/s writing into the host's memory. This limitation seems to be determined by the RIO2 and has been observed before [14]. Using a special configuration of the low-level library with pre-loaded data, i.e. suppressing the DMA and only having one transfer, a *bandwidth* of 96 Mbyte/s has been measured. Future versions of the Systran cards will have a dual-port memory which will overcome the first limitation and increase the throughput.

The *overhead* values are very similar for all three classes and it can be concluded that the round-trip-time and the overhead due to a class 1 connection cannot be measured this way. The end-to-end performance for messages of 1 kbyte is 23 Mbyte/s or 24 kHz, and for messages of 10 kbyte 45 Mbyte/s or 4.6 kHz.

3.2. Interphase cards on RIO2s

Fig. 3 shows the data rate versus the size of the messages between two Interphase cards on RIO2s in point-to-point configuration [15].

The connection-less class 3 transfer results in the highest data rates, whereas the connection-oriented classes 1 and 2 transfers have a lower data rate because they have to wait for ACKs (class 1) for R_RDYs and ACKs (class 2). The angles in the curves for classes 1 and 2 transfers that occur at the frame size (= 2 kbyte) can be explained by the credit-based flow-control mechanism which allows pipelining of the R_RDYs and ACKs in multi-frame sequences, so that transmission of the next frame can be stared before the ACK (R_RDY) of the previous frame has been received.[4] The degree of pipelining depends on the BB credit (for R_RDYs) and EE credit (for ACKs). A value of greater or equal 3 was found to lead to the highest data rates. Alternatively, the ACK_0 mechanism could be used. In this mechanism only one ACK at the end of a sequence is required. The border between SFS and MFS is defined by the programmable frame size and it has been found that reducing the frame size can increase the data rate [13]. The final value has to be chosen depending on the kind of traffic expected.

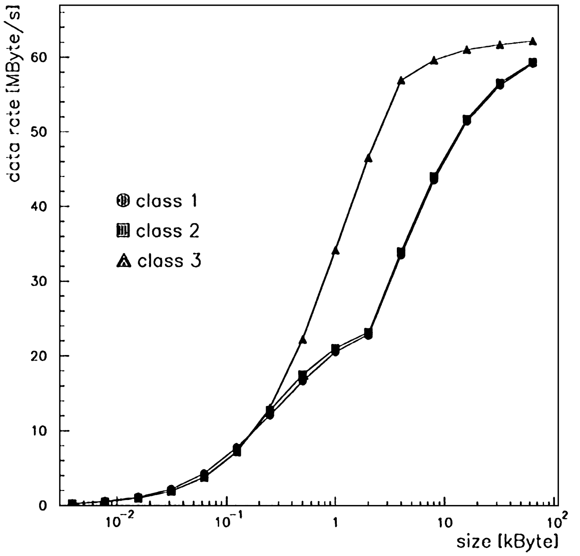

Fig. 3. Data rate for the Interphase cards.

The fits for classes 1 and 2 have been carried out for the SFS and MFS separately: for SFS the *overhead* values are 14 μs and the *bandwidth* 31 Mbyte/s; for MFS the *overhead* values are 53 μs and the *bandwidth* 63 Mbyte/s. For class 3 the *overhead* value is 6 μs and the bandwidth 63 Mbyte/s, the same as for classes 1 and 2 for MFS. The saturation bandwidth is 63% of the FC capacity and is limited by the PCI as can be seen using the analyser (see Section 3.3). The *overhead* values cannot be used to measure the round-trip-time and

[4] This effect has not been so pronounced with the Systran cards due to the lower rates (see Fig. 2)

the class 1 connection overhead. The end-to-end performance for data messages of 1 kbyte is 34 Mbyte/s or 35 kHz for class 3 and 21 Mbyte/s or 22 kHz for classes 1 and 2. For messages of 10 kbyte the end-to-end performance is 60 Mbyte/s or 6.1 kHz for class 3 and 47 Mbyte/s or 4.8 kHz for classes 1 and 2.

3.3. *Analyser results*

The analyser has been used to measure some parameters which could not be measured with the statistical method. The propagation delay on the optical fibres could be determined to be 4.9 ± 0.1 ns/m, corresponding to a refraction index of $n \approx 1.47$. On the Interphase cards it was seen that the *bandwidth* for reading from PCI is about 66 Mbyte/s, while that for writing to PCI is about 76 Mbyte/s. On the Systran cards, read and write bandwidth of the Tachyon chip from and to the local memory of more than 100 Mbyte/s have been measured, when there was no contention of the local memory bus. Other specific parameters of the Tachyon chip could be measured: the R_RDY response time of the Tachyon chip after receiving a frame is 560 ± 10 ns, the time to process a R_RDY when waiting for BB credit and the data are available is 430 ± 10 ns. The overhead due to establishing a class 1 connection is 3.0 ± 0.2 μs.

3.4. *Interoperability*

The tests have been repeated with the two different types of the FC cards being connected in a point-to-point configuration. It turns out that the cards from the two different manufacturers interoperate without any problems running on RIO2s and using the low-level library.

4. Performance in arbitrated loop configuration

4.1. *Topology*

In an FC arbitrated loop several nodes share the same medium. Only one node can transmit data at a time. A node which wants to transmit data must arbitrate for the loop control using the ARB primitive. Once it obtains the control, it uses the OPN primitive to open a connection to the destination node on the loop. The transmission of data then follows the point-to-point transfer. When all data are sent the node uses the CLS primitive to close the connection and to release the loop control for another phase of arbitration. A typical arbitrated loop configuration with four nodes is shown in Fig. 4.

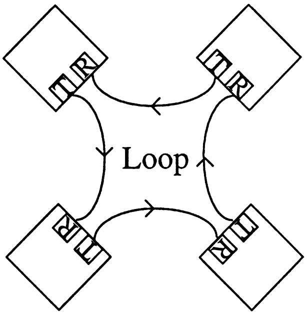

Fig. 4. Arbitrated loop with four nodes.

The loop initialization protocol is used to dynamically assign arbitrated loop physical addresses (AL_PA). A unique world-wide name (WWN) for each node is used to distinguish between the nodes. It is task of the user or system software to map between static WWNs and dynamic AL_PAs. The Tachyon chip supports only classes 2 and 3 on the arbitrated loop.

4.2. *Loop overhead*

The overhead of the transfers due to the arbitration and opening phase compared to simple point-to-point transfers has been measured on RIO2s and with the low-level library [16]. It is of the order of a few percent for most message sizes, except for message sizes of a tenth of the frame size where the degradation is of the order of 10%.

Using the analyser it was found that every node introduces a delay of 225 ± 10 ns to forward the ARB and OPN primitives. Since the loop's BB credit at opening was chosen to be 0, an additional time of 510 ± 10 ns has to be added for the generation and processing for the first R_RDY. With two nodes on the loop, the overhead is about 2 μs. The total *overhead* for a whole sequence depends on how often a connection has to be opened. If the Tachyon chip is programmed to close the loop when it has no more data or BB credit available (see Section 4.4) the loop will have to be re-opened once the next frame(s) can be sent. The total *overhead* increases with the message size, effectively reducing the *bandwidth*.

4.3. FXLP with PCs

Tests with Systran cards on PCs have been performed [17]. DMA measurements of data transfers between the PC's memory and the memory of the FC/PMC card have been carried out: the *overhead* is 50 μs, the *bandwidth* from the FC/PMC card to the PC's memory is 85 Mbyte/s, and 100 Mbyte/s in the other direction. Using the FXLP driver for Windows NT from Systran, the end-to-end performance of data transfers between two PCs was measured to have an *overhead* of 75 μs and a *bandwidth* of 40 Mbyte/s.

In order to interoperate FC cards on PCs and RIO2s, the FXLP driver for Windows NT has been used on the PCs and the FXLP extension of the low-level library on the RIO2s. Sending messages of 64 kbyte from PC to RIO2, a sustained data rate of 40 Mbyte/s has been observed. From RIO2 to PC the data rate was 33 Mbyte/s. Sequential send had to be used in the low-level library as it turns out that the FC BB credit mechanism to slow down the sender works correctly, but that the FXLP driver cannot handle receiving data faster than 40 Mbyte/s. PCs and RIO2s can interoperate using the existing FXLP driver and the low-level library in a loop configuration.

4.4. Overlapping transfers

On an arbitrated loop several transfers of MFS can overlap, if the Tachyon chip is programmed to close the loop when it has no more data or BB credit available. Another transfer can make use of the inter-frame gaps and thus efficiently overlap with the first transfer. When overlapping transfers and using class 3, frames can be lost when the transfers go to the same receiving node. The Tachyon chip can only handle one incoming MFS at a time and will drop all incoming class 3 frames not belonging to the current sequence. Using class 2, the Tachyon chip also drops incoming class 2 frames not belonging to the current sequence, but in addition, it will send back a frame (P_BSY) to indicate that it is busy. The sending node can then retry to send the frame. The Tachyon chip can be programmed to wait for the ACK to the first frame and to retry up to 16 times if an P_BSY is received. When 16 retries are not sufficient, the user level of the software has to take action. Of course, the penalty of this method is a bigger overhead for the transfers due to the blocking on the first frame, but the advantage is that no frames can be lost.

If the Tachyon chip was not programmed to close the loop when it has no frames to send, then transfers could not overlap and an EB-like application on an arbitrated loop would be limited to the bandwidth of a single transfer. Allowing transfers to overlap and using class 2 with blocking on the first frame and retry, an EB-like application on the arbitrated loop can increase the performance and not lose any frames. This method has been tested with two nodes acting as EB sources and two nodes acting as EB destination. The data rate versus message size is shown in Fig. 5. The aggregated data rate does not scale with the number of destinations since the maximum bandwidth of this configuration is limited to 100 Mbyte/s and the nodes spend additional time for the arbitration. However, an increase of about 50% between the 1 × 1 EB-like system to the 2 × 2 EB-like system can be observed and a total bandwidth value of 80 Mbyte/s has been fitted in the latter case, corresponding to 80% of the maximum bandwidth. This value is, however, not reached for massage sizes of 64 kbyte due to an increased *overhead* of 40 μs and will only be reached at message sizes of about a few hundred kbyte.

4.5. Broadcast

Broadcast on an arbitrated loop is defined for class 3 and for the well known AL_PA = 0xFF. A loop credit has to be given for as many frames which have to be broadcast since no R_RDY is sent back from the nodes. Broadcast of SFS and MFS of up to 4 kbyte have been tested. The data were correctly received by all the nodes on the loop,

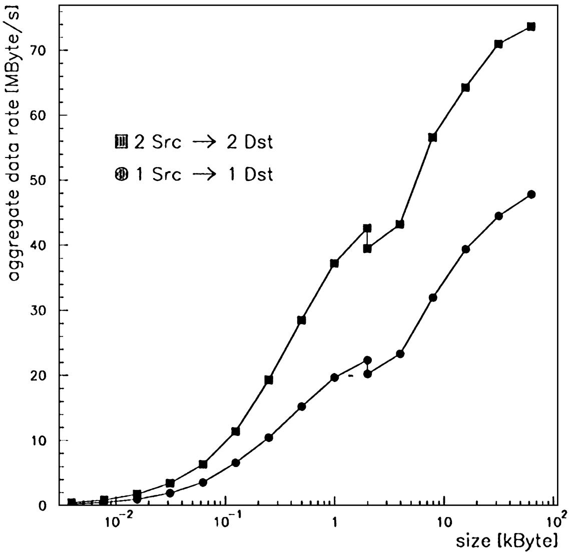

Fig. 5. Event building using Arbitrated Loop (Class 2).

including the sender. The performance of broadcast is the same as for unicast class 3 transfers. Multicast is not defined on a loop.

5. Performance in fabric configuration

5.1. Brocade switch

The FC cards have been tested successfully with the *SilkWorm* FC switch from Brocade Communication System, Inc. [18,19]. This switch supports only classes 2 and 3 and allows full connectivity between any of its ports. A typical configuration with four nodes is shown in Fig. 6.

The switch requires an explicit login of all ports at initialization using an FLOGI frame. A physical address is assigned by the switch corresponding to the port the node is connected to.

5.2. Switch overhead

For class 2 comparisons of point-to-point transfers and single transfers going through the switch have shown the delay of the sequences through the switch to be of 2 ± 1 μs, confirming the manufacturer's specification. Using the analyser, the same time, measured as the time between a class 2 or class 3 start-of-frame going into the switch and coming out of it was found to be 1740 ± 10 ns. No limit on the throughput was observed even when using the special configuration of the low-level library for the Systran cards (see Section 3.1) which transfers data at 96 Mbyte/s.

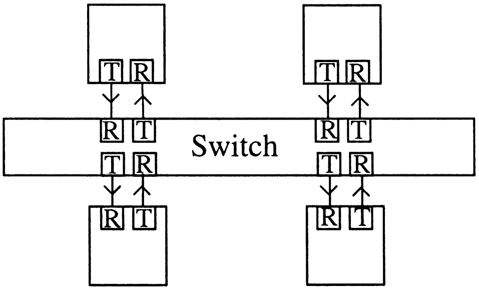

Fig. 6. Switch with four nodes.

The switch responds faster to incoming frames than the Tachyon chip. The R_RDY is generated 430 ± 10 ns after the end-of-frame. When there is a frame in the switch which cannot be sent due to missing BB credit the time to process an R_RDY takes between 500 and 1200 ns due to time-slicing in the switch [18].

5.3. Concurrent transfers

In order to test the switch's behaviour under load, comparisons of single transfers and concurrent, but mutually exclusive transfers, using classes 2 and 3 have been carried out. No degradation of the rates has been observed. Up to three concurrent transfers have been shown to have scaling behaviour: the data rates of the transfers add up and no internal bottle-neck was encountered. The switch has been loaded with about 200 000 frames/s (for messages of 4 byte) and with about 180 Mbytes/s (for messages of 64 kbyte). This is only a small portion of the load the switch can accept, the manufacturer states that the switch is capable of accepting a total of 8 million frames/s and 3.2 Gbyte/s aggregate bandwidth.

An EB-like application can use this scaling of the rates directly if it has an external mechanism of synchronisation guaranteeing that all transfers at

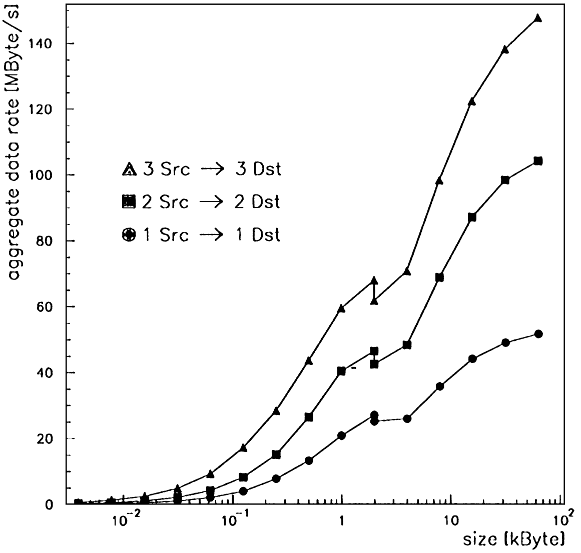

Fig. 7. Event Building using an FC Switch (Class 2).

any given time are mutually exclusive. If such a mechanism is not available several MFS can overlap at an EB destination. Due to the limitation of the Tachyon chip, the method described in Section 4.4, using class 2 transfers, blocking on the first frame and retry has to be used to guarantee that no frames are lost. This method has been tested with up to six FC cards from Systran and Interphase. The data rates are shown in Fig. 7, for 1 × 1, 2 × 2 and 3 × 3 EB-like systems. EB using the switch shows very clearly scaling behaviour of the data rate with the number of destinations. In particular, the data rate for message sizes of 1 kbyte/s scales with 20 Mbyte/s and for 10 kbyte with 35 Mbyte/s for each destination. Comparing these values to the values for point-to-point configuration (see Sections 3.1 and 3.2), there is no difference for messages of 1 kbyte (SFS). The throughput decreases by about 30% for messages of 10 kbyte (MFS) because of the bigger *overhead* due to the blocking on the first frame.

5.4. Multicast and broadcast

The *SilkWorm* switch allows multicast and broadcast. Multicast which is only defined for class 3 requires registration of nodes in a multicast group. The registration is handled by an alias server which is an integral part of the SilkWorm switch. Broadcast is also only defined for class 3 but does not require registration: it uses the well-known address identifier 0xFFFFFF. The protocol to register multicast groups with the alias server of the switch has been implemented and multicast and broadcast of transfers have been tested successfully. The throughput, however, decreases by an order of magnitude compared to unicast class 3 transfers. This degradation is not fully understood. The manufacturer did not publish how the routing mechanism works and how the routing information from the alias server is obtained.

5.5. Network management

The switch can be configured and monitored using a dedicated ethernet port and the telnet protocol. The switch also contains an SNMP server and defines the standard MIBs as well as the experimental FC MIB and a switch specific MIB. The name server is an integral part of the Brocade switch and allows the mapping of physical addresses to world-wide names. The protocol to read the information of the name server has been implemented in the low-level library and can be used to obtain a complete map of the nodes connected to the switch. It has to be noted that the standard defining the name server (FC-GS2) is not yet stable and that FC does not define an address resolution protocol for the mapping of the logical world-wide names to the FC physical addresses. The name server is intended to overcome this shortcoming.

5.6. Cascaded switches and attached loops

Other interesting features of the switch which have not been tested in our laboratory are the cascading of switches and the attaching of arbitrated loops to the switch. Both features allow to build big networks for hundreds of nodes. Attached loops will be available soon and allow to build EB systems in which arbitrated loops are used to group several EB sources sharing the bandwidth of the loop before they are connected to a network of switches. The number of switch ports can thus be reduced and each switch port will be used much

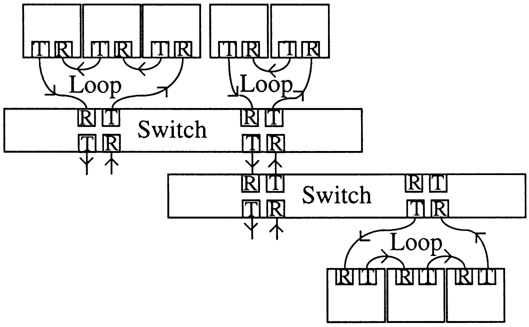

Fig. 8. Network using arbitrated loops and cascaded switches.

more efficiently. This will reduce the cost of the system considerably. A typical configuration using attached loops and cascaded switches is shown in Fig. 8.

The scaling capabilities of cascaded switches and the behaviour of the end-to-end latency will have to be investigated. The behaviour of huge networks carrying the data traffic for the ATLAS EB system will have to be simulated. Previous simulations, however, show already that if the individual switch latency is small compared to the total transfer time and if the switch is non-blocking up to the rates of the ATLAS experiment, that a cascaded switching network will be scaling [20]. These simulations will have to be continued once results of measurements with cascaded switches and attached loops are available.

6. Conclusion

Experience with FC technology has been obtained. Hardware and software components have been tested successfully with simplified network applications in the environment of the ATLAS DAQ prototype " − 1". Measurements have shown sustained data rates of 50–60 Mbyte/s from end-user to end-user. Low latencies of 10–20 μs, including the software overhead, have been observed. The equipment has been tested in several network topologies, like point-to-point, shared medium and switched topologies. FC offers potentially the high bandwidth and the high degree of connectivity required for the ATLAS EB system (see Table 1). In addition, FC could also be used for other applications in the ATLAS DAQ system, e.g. local data collection and mass storage.

Using the FC equipment for EB, the following two features have to be considered: a limitation in the assembly of FC sequences and inefficiency in the interfaces between the FC card and the host system. The Tachyon chip can only assemble one MFS at a time. If event fragments are greater than one frame (of up to 2 kbyte), an additional synchronization mechanism (see Section 4.4) has been proposed. This mechanism provides reliable data transfer on arbitrated loops and with an FC switch, at the cost of a performance penalty of about 30%. It has to be investigated if eventually FC protocol chips will be available which are able to support multiple concurrent MFS. Alternatively, it could be investigated if class 1 transfers could be used for EB, in particular with switches which support stacked connect-requests.

The interface between the FC card and the host system has a hardware and a software aspect. The host system's I/O bus and the internal resources of the FC cards are not matched to the bandwidth of the FC link. Effectively, only 50–60% of the FC bandwidth is available to the end-user. Commercial drivers for the platforms in the environment of the ATLAS DAQ prototype " − 1" do not make efficient use of the FC hardware. The driver for Windows NT does not support a switch topology and in order to interoperate FC cards on PCs and RIO2s the loop configuration has to be used. The only system software available for LynxOS which makes satisfactory use of the FC cards had to be developed especially for this purpose.

FC is a candidate technology for EB and will be used for an implementation of the EB in the ATLAS DAQ prototype " − 1" where issues like detailed integration and scalability will be addressed. The rapidly moving industrial developments will have to be watched in order to understand how FC compares to other technologies in terms of cost and performance as well as in terms of availability of components and software off-the-shelf and the longevity of this technology.

Acknowledgements

We would like to thank the people from Systran and from Brocade for their technical support and for temporary loan of equipment used in this evaluation. We are grateful to A. Razc, D. Samyn and T. Ladzinski from the CMS collaboration at CERN for the useful discussions and for lending us some FC/PCI cards needed to complete our tests.

References

[1] ATLAS Collaboration, Technical Proposal for a General-purpose pp Experiment at the Large Hadron Collider at CERN, CERN/LHCC/94-43, see http://atlasinfo.cern.ch:80/Atlas/Welcome.html.

[2] G. Ambrosini et al., The ATLAS DAQ and Event Filter Prototype " − 1" Project, presented at Computing in High-energy Physics 1997, Berlin, Germany, http://atddoc.cern.ch/Atlas/Conferences/CHEP/ID388/ID388.ps.

[3] For the different documents within the family of the Fibre Channel standard see the Fibre Channel Association, http://www.amdahl.com/ext/CARP/FCA/FCA.html.

[4] Systran Corp., http://www.systran.com.

[5] Interphase Corp., http://www.iphase.com.

[6] Tachyon, HP-FC-5000, Hewlett-Packard Comp., http://tachyon.rose.hp.com.

[7] Technobox Corp., http://www.technobox.com.

[8] RIO2 8060, RIO2 8061 and RTPC 8067 are products from Creative Electronics Systems S.A., http://www.ces.ch/CES_info.

[9] IBM Corp., http://www.chips.ibm.com/products/embedded/chips/ibm82660.html.

[10] Lynx Real-time Systems, Inc., http://www.lynx.com.

[11] R. Spiwoks, The Fibre Channel Library in the ATLAS DAQ Prototype, ATLAS DAQ Prototype Note 083, http://atddoc.cern.ch/Atlas/postscript/Note083.ps.

[12] KFKI Research Institute for Particle and Nuclear Physics. http://www.rmki.kfki.hu/detector/preproc/pp-broch.html.

[13] R. Spiwoks et al., Evaluation of FC/PMC Cards in the Environment of the ATLAS DAQ Prototype, ATLAS DAQ Prototype Note 021, http://atddoc.cern.ch/Atlas/postscript/Note021.ps.

[14] M. Joos et al., An Evaluation of VMEbus POWERPC-based Processor Boards Running LynxOS Operating ystem, http://www.cern.ch/ECP-ESS/Reports/PPC_Evaluation/Draft_1.1.3/Title.html.

[15] R. Spiwoks, Performance Tests with FC/PCI Cards in the Environment of the ATLAS DAQ Prototype, ATLAS DAQ Prototype Note 034, http://atddoc.cern.ch/Atlas/postscript/Note034.ps.

[16] R. Spiwoks, Experience with FC Arbitrated Loop in the Environment of the ATLAS DAQ Prototype, ATLAS DAQ Prototype Note 036, http://atddoc.cern.ch/Atlas/postscript/Note036.ps.

[17] M. Romano, Evaluation of FC/PMC Cards on PCs, ATLAS DAQ Prototype Note 057, http://atddoc.cern.ch/Atlas/postscript/Note057.ps.

[18] Brocade SilkWorm Switch, Brocade Communications Systems, Inc., see http://www.brocadecom.com.

[19] R. Spiwoks, E. v.d. Bij, Evaluation of a FC Switch in the Environment of the ATLAS DAQ Prototype Note 053, http://atddoc.cern.ch/Atlas/postscript/Note053.ps.

[20] R. Spiwoks, Evaluation and Simulation of Event Building Techniques for a Detector at the LHC, Ph.D. Thesis, University of Dortmund, Germany, 1995; CERN/THESIS/96-002; http://atddoc.cern.ch/ ~ spiwoks/papers/diss/diss.ps.